Assignment II: Machine Learning Tasks on Preprocessed Education Dataset

CSE 303 - Statistics for Data Science

Submission Deadline: 14.01.2025

Objective

This assignment focuses on building foundational skills in applying machine learning techniques to the preprocessed medical dataset. Students will work on tasks such as data preprocessing, model selection, training, evaluation, and feature importance analysis.

Dataset

Use the education dataset provided in Assignment I after performing data cleaning, feature engineering, and other preprocessing steps.

Tasks

1. Data Preprocessing (15 points)

- Ensure the dataset is free from missing values and outliers (use preprocessed dataset from Assignment I).
- Perform label encoding or one-hot encoding for categorical columns.
- Standardize or normalize numerical columns for better model performance.
- Create train-test split (70%-30%).

Deliverables:

- Preprocessed dataset ready for machine learning tasks.
- Explanation of preprocessing steps with accompanying code.

2. Exploratory Data Analysis for Model Selection (10 points)

- Analyze correlations and visualize relationships between features and the target variable.
- Identify potential predictive features for the target variable.

Visualization:

- Correlation heatmap.
- Scatter plots or bar charts for key relationships.

3. Model Training and Evaluation (30 points)

- Train three different machine learning models (e.g., Logistic Regression, Random Forest, XGBoost).
- Evaluate models using appropriate metrics: Accuracy, Precision, Recall, F1 Score, and ROC-AUC.

Deliverables:

- Model performance comparison table.
- ROC curves for all models.

4. Hyperparameter Tuning (15 points)

- Select the best-performing model and tune its hyperparameters using grid search or random search.
- Explain the tuning process and parameters selected.

Deliverables:

- Best hyperparameters and corresponding model performance.

5. Feature Importance Analysis (15 points)

- Use the best-performing model to analyze feature importance.
- Identify and discuss the top 5 features contributing to model predictions.

Visualization:

- Bar chart for feature importance.

6. Model Interpretability (15 points)

- Use SHAP or LIME to explain predictions of the best-performing model.
- Discuss the explainability insights gained.

Deliverables:

- SHAP or LIME plots with an explanation.

Submission Guidelines

- Complete the assignment in a **Google Colab notebook**.
- Share the notebook link with **view-only permissions**.
- Document each step using comments and markdown cells.
- Provide detailed justifications for your approach and decisions.

Evaluation Criteria

The assignment will be evaluated through a **viva** on **12.01.2025**, where students must justify their choices and explain the insights gained.

Good luck!